*Jurnal Evaluasi Pendidikan Vol. 2 No. 2, Oktober 2011, 132-146*
*http://doi.org/10.21009/JEP*

# *INAPPROPRIATNESS SCORE BASED ON ITEM RESPONSE THEORY*

**Budi Santoso**
Fakultas Keguruan dan Ilmu Pendidikan, Universitas Sriwijaya
Jl. Raya Palembang-Prabumulih KM. 32 Indralaya, Ogan Ilir
yayasanbudi_s@yahoo.com

*DOI: doi.org/10.21009/JEP.022.02*

***Abstract***

*The objective of this research is to compare inappropriateness based on Item response theory as an effect of using model of scoring and number of options for multiple choice items. The research method used quasi experiment. Independent variable of this research are Correct Score (CS), punishment score (PS), and Reward Score (RS). And also three and five option multiple choice items. To see this comparison, the proportions of fair score are compared by using $\ell_{gz}$ index. For the student corrected by CS model compared between the group who answer three choice item and group who answer five choice item. While for group of student doing test with a five choice item compared between the couple of CS, PS and RS. The result of this research indicate that for low ability ($\theta <- 1$) student, the use of CS model for student who answer a five choice item compared has more fair score than students who answer a three choice item. But for high ability ($\theta > 1$) students there are no difference fair score among the two groups. For the students who answer a five choice item, indicate that for high ability student, the use of PS and RS model have more fair score compare with CS model, but for PS model does not yield fair score difference compare with RS model. Meanwhile for low ability student there is no difference in fair score among the use of three models of scoring.*

***Keywords:*** *inappropriateness, correct score, punishment score, reward score, multiple choice item*

# KETIDAKWAJARAN SKOR BERDASARKAN TEORI RESPONSI BUTIR

**Budi Santoso**
Fakultas Keguruan dan Ilmu Pendidikan, Universitas Sriwijaya
Jl. Raya Palembang-Prabumulih KM. 32 Indralaya, Ogan Ilir
yayasanbudi_s@yahoo.com

**Abstrak**

Tujuan penelitian ini adalah untuk membandingkan ketidakwajaran skor berdasarkan Teori Responsi Butir ditinjau dari penggunaaan model penskoran dan jumlah pilihan jawaban pada tes pilihan ganda. Metode penelitian yang digunakan adalah *quasi experiment*. Variabel independen penelitian ini adalah model penskoran yang terdiri atas model penskoran jawaban benar (*correct score*), penskoran hukuman (*punishment score*), dan penskoran hadiah (*reward score*). Variabel independen lainnya adalah tes pilihan ganda dengan tiga pilihan dan lima pilihan jawaban. Untuk melihat perbandingan ini, peneliti membandingkan proporsi skor wajar yang dihitung menggunakan indeks kewajaran $\ell_{gz}$. Untuk siswa yang dikoreksi menggunakan model CS dibandingkan antara siswa yang menjawab tes dengan lima pilihan dan tiga pilihan. Untuk siswa yang mengerjakan tes lima pilihan dibandingan antara kelompok yang dikoreksi menggu-nakan CS, PS dan RS secara berpasangan. Hasil penelitian menunjukkan bahwa untuk siswa dengan kemampuan rendah ($\theta <- 1$), penggunaan model CS untuk siswa yang menjawab tes lima pilihan mempunyai skor yang lebih wajar dibandingkan dengan siswa yang mengerjakan tes tiga pilihan. Tetapi untuk siswa dengan kemampuan tinggi ($\theta > 1$) menunjukkan tidak ada perbedaan skor wajar diantara dua kelompok tersebut. Untuk siswa yang mejawab tes lima pilihan, hasil penelitian menunjukkan bahwa untuk siswa dengan kemampuan tinggi, siswa yang dikoreksi menggunaan model penskoran PS dan RS memiliki skor yang lebih wajar dibandingkan siswa yang dikoreksi menggunakan model CS. Tetapi untuk penskoran PS dibandingkan penskoran RS menghasilkan proporsi skor wajar yang tidak berbeda. Untuk siswa dengan kemampuan rendah ketiga model penskoran menghasilkan skor wajar yang tidak berbeda diantara ketiga model penskoran tersebut.

**Kata kunci:** ketidakwajaran, penskoran jawaban benar, penskoran hukuman, penskoran hadiah, tes pilihan ganda tiga

## PENDAHULUAN

Salah satu kegiatan penting untuk meningkatkan mutu pendidikan di sekolah adalah evaluasi atau penilaian. Salah satu proses pelaksanaan penilaian adalah menggunakan tes. Tes yang dimaksud adalah suatu upaya untuk melakukan pengukuran terhadap tingkat pencapaian atau hasil belajar peserta didik. Tes diartikan sebagai alat untuk mengukur pengetahuan atau penguasaan obyek ukur terhadap seperangkat konten atau materi tertentu. Wiersma dan Jurs (1990: 8) menyatakan bahwa tes mempunyai arti yang sangat dekat dengan

pengukuran dan asesmen. Tes umumnya merujuk kepada sekumpulan butir pertanyaan yang didesain untuk diberikan kepada satu atau lebih siswa di bawah kondisi tertentu. Nitko (2001: 5) mendefinisikan tes sebagai suatu instrumen atau prosedur yang sistematis untuk mengobservasi dan menggambarkan satu atau lebih karakteristik dari seorang siswa.

Salah satu bentuk tes hasil belajar tipe objektif yang paling banyak dikenal dan digunakan adalah tes obyektif berbentuk pilihan ganda. Bentuk tes ini dapat mengukur lebih efektif dibandingkan bentuk jawaban singkat dan bentuk betul-salah maupun bentuk menjodohkan. Di samping itu, bentuk tes ini juga dapat mengukur variasi hasil belajar yang kompleks dalam area pengetahuan, pemahaman, dan penerapan (Gronlund, 1990: 39). Bahkan Ebel dan Frisbie (1991: 154) menyatakan bahwa hampir semua kemampuan yang dapat dites menggunakan bentuk tes objektif lain seperti jawaban singkat, melengkapi, benar-salah, menjodohkan atau esei dapat diukur menggunakan bentuk tes pilihan ganda. Sebuah tes pilihan ganda mengandung suatu stem (pernyataan, pertanyaan, ungkapan, atau sejenisnya) dan beberapa pilihan (mulai dari dua pilihan sampai lima pilihan) diantaranya hanya satu jawaban yang benar (Aiken, 1997: 466).

Azwar (1987: 75) mengatakan bahwa tes bentuk pilihan ganda mempunyai keunggulan dan kelemahan. Menurutnya, tes bentuk pilihan ganda mempunyai keunggulan karena (1) dalam waktu tes yang singkat dapat memuat banyak soal, (2) pemeriksaan jawaban dan pemberian skor mudah serta cepat, (3) penggunaan lembar jawaban, menjadikan tes efisien dan hemat biaya, (4) kualitas soal dapat dianalisis secara empirik, (5) obyektivitas tinggi, (6) umumnya mempunyai reliabilitas yang memuaskan. Sedangkan kelemahan tes pilihan ganda adalah: (1) pembuatannya sulit, memakan waktu dan tenaga, (2) tidak mudah untuk mengungkapkan kompetensi tinggi, dan (3) ada kemungkinan jawaban benar semata-mata karena tebakan.

Dalam menentukan jumlah pilihan yang digunakan pada soal pilihan ganda, Gronlund dan Linn (1990: 181) berpendapat bahwa tidak ada bilangan yang paling tepat tentang jumlah pilihan yang harus digunakan. Biasanya berkisar mulai dari tiga, empat atau lima pilihan yang sering digunakan. Makin banyak pilihan makin kecil peluang mendapatkan jawaban benar dengan cara menebak. Para peserta tes seringkali melakukan tebakan dalam tes pilihan ganda. Tebakan ini dapat berupa tebakan sebarang (*blindly guessing*) atau tebakan yang menggunakan *partial information* (Frary, 1980: 80). Berkaitan dengan ini Nunnaly (1970: 176) menyatakan peserta tes seringkali melakukan teknik *guessing* dengan melakukan eliminasi terhadap pilihan jawaban yang dianggap tidak mungkin benar. Oleh karena itu, jumlah pilihan sesungguhnya cenderung lebih sedikit dibanding jumlah pilihan yang diberikan sehingga efek *guessing* menjadi berkurang.

Tidak ada aturan baku mengenai berapa jumlah pilihan yang sesuai. Jumlah pilihan dalam soal pilihan ganda yang biasa digunakan adalah mulai dari

dua pilihan sampai dengan lima pilihan. Bila dilihat dari segi efisiensi dan dari reliabilitas tes. Menurut Lord (1980: 106) jumlah optimal pilihan dalam tes pilihan ganda adalah tiga. Ia mengatakan bahwa jika penambahan jumlah pilihan tidak menambah waktu dan biaya, dari pertimbangan umum nampaknya bahwa makin banyak pilihan makin baik. Alasan matematis tentang hal ini diberikan oleh Tversky yang juga menyimpulkan bahwa tiga pilihan jawaban tersebut adalah optimal.

Setelah peserta tes mengerjakan tes pilihan ganda, selanjutnya dilakukan penskoran terhadap respons yang diberikan. Skor tersebut merupakan informasi hasil tes yang merupakan representasi dari kemampuan siswa. Untuk melakukan penskoran ini dapat digunakan beberapa alternatif model penskoran. Ada tiga model penskoran yang dikenal pada tes pilihan ganda antara lain: 1) penskoran dengan menghitung jumlah jawaban yang benar saja (*correct score*) sering juga disebut *number right score* atau *conventional scoring*, 2) penskoran dengan memberi sanksi pada jawaban yang salah *(punishment score)* atau *rights minus wrongs correction*, dan 3) penskoran dengan memberi hadiah pada butir yang tidak dijawab (*reward score)* atau *correcting row score*, kedua model terakhir ini sering juga disebut sebagai *formula scoring* (Crocker and Algina, 1986: 401). Penerapan model penskoran yang berbeda-beda dapat berdampak pada skor yang diperoleh masing-masing peserta karena peserta akan mempertimbangkan kemungkinan untuk menjawab dengan cara menebak atau tidak menjawab butir soal yang sedang dikerjakannya dan pada gilirannya dapat menyebabkan terjadinya ketidakwajaran skor.

Model penskoran tes pilihan ganda dewasa ini yang cenderung digunakan adalah menjumlahkan skor jawaban yang benar saja (*correct score = CS*) sebagai skor peserta tes dimana butir yang benar diberi skor 1 dan yang salah atau tidak diisi diberi skor 0. Model penskoran seperti ini akan menyebabkan peserta tes berspekulasi dalam menjawab tes. Sebagai skor prestasi, model penskoran ini selain memberi peluang peserta melakukan *blindly guessing* berimplikasi pula pada tidak atau kurangnya validitas tes tersebut serta menurunnya tingkat indeks reliabilitas tes. Dalam hal ini Hopkins & Antes (1985: 213) mengungkapkan bahwa *guessing* dalam tes pilihan ganda dapat menurunkan nilai validitas butir dan reliabilitas tes. Selain itu, skor pencapaian peserta tes yang dicapai secara murni karena peserta mengetahui pilihan jawaban yang benar dan peserta yang menjawab dengan *guessing* juga sulit dibedakan bila menggunakan penskoran *correct score*. Bila dikaitkan dengan hasil penskoran hasil suatu tes pilihan ganda dengan butir-butir soal yang dibiarkan tidak dijawab (*omit*) oleh peserta tes, tentu akan lain pencapaian skornya. Demikian pula, bila penskoran tersebut dikaitkan dengan banyaknya pilihan jawaban yang disediakan.

Adanya kelemahan-kelemahan pada model *correct score* memunculkan alternatif model penskoran untuk mengurangi *guessing* yaitu model penskoran hukuman *(punishment score = PS*) yang merupakan model koreksi terhadap *guessing*. Model ini didasarkan pada dua asumsi. *Pertama*, semua tebakan

bersifat buta, masing-masing pilihan memiliki peluang yang sama untuk dipilih. Asumsi ini menolak kemungkinan bahwa seorang peserta tes mungkin melakukan tebakan menggunakan teknik *partial information* yang menyebabkan peluang menjawab benar menjadi lebih besar dari 1/k (Allen & Yen, 1979: 145). Bila model ini diterapkan pada kondisi ini, maka skor terkoreksi yang diperoleh men-jadi *underestimate*. *Kedua*, setiap jawaban yang salah diperoleh hasil tebakan. Seorang peserta tes dianggap tidak pernah merespon berdasarkan kesalahan informasi atau kesalahan menuliskan jawaban pada lembar jawaban. Asumsi ini mengabaikan kemungkinan menjawab salah yang disebabkan karena kesalahan memahami soal (Thorndike, 2005: 324).

Model *punishment score* merupakan model penskoran yang memperhitungkan jawaban salah yang direspon oleh peserta tes dengan jalan memberi hukuman dalam bentuk pengurangan skor. Algina menyebutnya juga dengan nama *right minus wrongs correction*. Asumsi dasar dari penggunaan rumus ini adalah jawaban salah merupakan hasil tebakan, sehingga sehingga jumlah yang salah dibagi dengan k – 1 merupakan hukuman bagi peserta tes yang menjawab dengan tebakan.

Pendekatan lain untuk mengatasi kelemahan *correct score* yakni model penskoran yang memperhitungkan butir yang tidak diisi atau dikosongkan (*omit*) oleh peserta tes dengan jalan memberi hadiah dalam bentuk tambahan skor. Model penskoran ini didasarkan pada pertimbangan tiga kemungkinan situasi: 1) peserta ujian mengetahui pilihan jawaban yang benar dan memilihnya, 2) peserta tidak memilih sama sekali pilihan jawaban yang disediakan, dan 3) peserta ujian menebak buta (*guessing blindly*) dan memilih satu dari k pilihan jawaban secara acak.

Hasil penskoran dalam suatu tes memberikan informasi atas kemampuan peserta tes yang diukur. Informasi hasil tes dapat saja tidak menjangkau sampai ke besaran atau dimensi yang hendak diukur oleh tes itu. Dapat pula terjadi hasil tes itu tercampur dengan besaran atau dimensi lain yang tidak dimaksudkan untuk diukur oleh tes itu sehingga hasil tes menjadi rancu. Di samping kedua tersebut, pelaksanaan tes yang kurang layak dapat menghasilkan informasi yang tidak benar yang tidak mencerminkan kemampuan peserta tes.

Pada pengujian semacam ini, skor yang diperoleh merupakan skor yang tidak benar atau timpang yang tidak mencerminkan kemampuan peserta tes sebenarnya. Jika sumber dari ketimpangan skor yang terjadi adalah dari peserta tes maka disebut ketidakwajaran skor. Hulin *et al.* (1983: 110) menamakan ketidakwajaran skor ini sebagai *inappropriatness* yakni bila seorang peserta tes dengan kemampuan tinggi salah dalam menjawab soal mudah dan peserta tes dengan kemampuan rendah banyak menjawab benar soal-soal yang sulit, seorang peserta tes tidak menjawab terlalu banyak soal yang mudah, atau seorang peserta tes menjawab secara acak keseluruhan tes.

Skor pada hasil tes ada kalanya tidak memberikan informasi yang benar tentang peserta tes. Dapat pula terjadi, hasil tes itu bercampur dengan besaran

atau dimensi lain yang tidak dimaksudkan untuk diukur dalam tes itu sehingga hasil tes menjadi bias. Selain itu, pelaksanaan tes juga dapat menjadi faktor yang menyebabkan tes menghasilkan informasi yang tidak benar. Pada pengujian semacam ini, skor yang diperoleh peserta tes merupakan skor yang tidak benar atau skor yang timpang. Ketimpangan skor pada ujian merupakan skor yang tidak memberikan informasi yang benar tentang hal yang dimaksud untuk diukur dalam ujian itu.

Dalam pengukuran pendidikan, ketimpangan skor yang mungkin terjadi perlu dicegah atau dihindari. Ketimpangan skor tersebut perlu dideteksi dan dari hasil deteksi ini diambil keputusan yang tepat tentang apa yang harus dilakukan terhadap hasil yang dicapai dalam pengukuran itu. Skor dicapai oleh peserta tes setelah mengerjakan sejumlah butir ujian, maka ada dua sumber tempat ketimpangan dapat terjadi masing-masing adalah peserta tes dan perangkat tes beserta butir-butirnya. Dengan demikian, ketimpangan skor pada tes dapat dipandang dari kombinasi antara letak ketimpangan dan bentangan ketimpangan. Ketimpangan itu dapat terjadi a) pada peserta tes atau butir tes dan b) secara individu atau secara kelompok.

Ketimpangan skor secara individu disebut sebagai ketidakwajaran skor. Ketidakwajaran skor terjadi apabila seorang peserta tes memperoleh skor yang tidak sesuai dengan kemampuan, padahal semua butir tes sudah baik. Peserta tes dapat saja memperoleh skor yang jauh lebih rendah dari skor yang seharusnya diperoleh berdasarkan kemampuannya, begitu juga sebaliknya. Hulin *et al.* (1983: 110) menamakan ketimpangan seperti ini sebagai *inappropriatness,* sedangkan Simanungkalit (1988: ii) menamakan ketimpangan ini sebagai ketidakwajaran. Ketimpangan skor dapat pula terjadi karena faktor individu, faktor individu tersebut antara lain: (1) gangguan emosional, sehingga mempengaruhi prestasi, (2) perasaan takut dengan situasi tes, sehingga menghasilkan respon yang tak normal (Gronlund and Linn, 1990: 76).

Pada teori responsi butir selain parameter taraf sukar butir (b) dan daya beda butir (a) terdapat parameter "c" yang disebut parameter "*pseudo chancelevel*" atau "*guessing parameter*" yang merupakan probabilitas menjawab benar dengan cara menebak. Crocker dan Algina (1986: 361) mengatakan bahwa untuk model tiga parameter dapat mengakomodasi tebakan dan tebakan harus dipertimbangkan sebagai suatu kemungkinan menjawab dalam tes berbentuk pilihan ganda. Apabila terjadinya ketidakwajaran skor itu tidak dapat dihindari, maka perlu dideteksi ketidakwajaran skor tersebut baik secara individu maupun secara kelompok. Hasil deteksi ini, dapat diambil keputusan yang tepat tentang apa yang harus dilakukan oleh penyelenggara tes terhadap hasil yang dicapai dalam pengukuran itu.

Di samping model penskoran, jumlah pilihan jawaban pada butir-butir tes secara tidak langsung dapat juga mempengaruhi hasil penskoran. Makin banyak jumlah pilihan yang disediakan maka semakin menyita perhatian peserta tes dalam mengerjakan tes tersebut, sehingga timbul pertanyaan seberapa besar hal

tersebut mempunyai pengaruh terhadap ketidakwajaran skor peserta tes. Secara teori probabilitas makin banyak pilihan jawaban makin kecil peluang siswa menjawab benar terutama bagi siswa yang menggunakan strategi menebak buta (*blindly guessing*). Siswa yang relatif pandai atau siswa yang tidak menggunakan strategi tebak buta, jumlah pilihan tidak menjadi permasalahan karena siswa menggunakan strategi menebak dengan memperhitungkan jawaban-jawaban yang pasti salah berdasarkan kemampuan yang dimiliki. Strategi ini disebut sebagai p*artial Information* yakni kemampuan mengurangi jawaban yang pasti salah dalam alternatif jawaban pada soal pilihan berganda (Frary, 1980: 80).

Berdasarkan uraian di atas, dirasakan cukup urgen untuk dilakukan penelitian terhadap variabel-variabel yang memungkinkan terjadinya ketidakwajaran skor pada model penskoran dan banyak option pada tes pilihan ganda.

**METODE PENELITIAN**

Penelitian ini merupakan penelitian kuasi eksperimen dengan variabel bebas model penskoran dan jumlah pilihan dalam soal pilihan ganda. Model penskoran yang digunakan terdiri atas tiga model penskoran yakni model penskoran *correct score* (CS), *punishment score* (PS) dan *reward score* (RS). Jumlah pilihan dalam tes pilihan ganda terdiri atas dua macam jumlah pilihan yakni tiga dan lima pilihan. Sedangkan variabel terikatnya adalah proporsi skor wajar berdasarkan indeks kewajaran skor $\ell_{gz}$. Indeks $\ell_{gz}$ dihitung berdasarkan fungsi karakteristik butir model logistik tiga parameter.

Indeks kewajaran $\ell_{gz}$ didapat melalui transformasi nilai baku dari indeks kewajaran $\ell_g$ yakni:

Indeks kewajaran geometrik baku $\ell_z$ (Naga, 2012: 612)

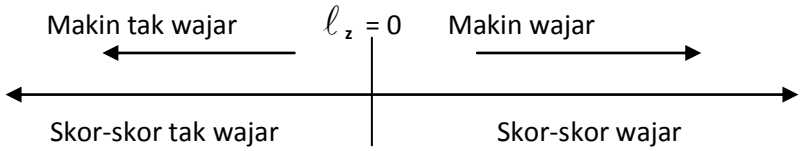
$$\ell_z = \frac{Lo - \mu_{\ell 0}}{\sigma_{\ell o}} \qquad (1)$$

dengan

$m_{lg}$ = nilai rata $\ell_o$ untuk semua peserta tes dan

$s\ell_g$ = simpangan baku $\ell_0$ untuk semua peserta tes

Nilai teoritis indeks kewajaran skor $\ell_z$ mempunyai rentang dari -∞ sampai ∞ dan displit menjadi dua kelompok yang terdiri dari indeks kewajaran negatif yang berasal dari skor-skor yang tidak wajar dan indeks kewajaran positif yang berasal dari skor-skor yang wajar. Secara teoritis indeks ini berkisar dari negatif tak hingga sampai positif tak hingga tetapi secara praktis indeks ini diperkirakan nilainya berkisar antara nilai -4 dan 4 sebagaimana distribusi normal baku.




**Gambar 1.** Distribusi Normal Baku

Rancangan penelitian yang digunakan adalah penelitian komparatif yakni membandingkan proporsi skor wajar berdasarkan indeks kewajaran yang diperoleh. Model rancangan ini dipilih karena indeks kewajaran yang digunakan adalah indeks kewajaran baku dengan rata-rata 0 sehingga yang dibandingkan adalah proporsi skor wajar yang mempunyai nilai indeks positif.

Penelitian ini dilakukan di beberapa SMA dengan responden siswa kelas X dengan sampel responden masing-masing 500 responden untuk masing-masing kelompok diambil dari 12 SMA dalam kota Palembang. Pengambilan sampel dengan teknik sampel purposif minimal 500 responden ini untuk tiap sel ini dengan tujuan untuk dapat mengestimasi parameter butir dan kemampuan, mengingat model yang dipakai adalah model karakteristik butir logistik tiga parameter (L3P). Persyaratan untuk estimasi parameter pada model L3P minimal menggunakan sampel sebanyak 500.

Untuk mendapat skor dan indeks kewajaran digunakan instrumen penelitian berupa perangkat tes matematika berbentuk pilihan ganda dengan tiga pilihan jawaban dan perangkat tes matematika pilihan ganda dengan lima pilihan jawaban dengan masing-masing perangkat tes terdiri atas 30 butir yang mewakili pokok bahasan matematika di semester genap kelas X SMA. Sebelum instrumen digunakan, dilakukan validasi panel terlebih dahulu dan dilakukan ujicoba untuk perbaikan instrumen dan validasi serta perhitungan reliabilitas instrumen. Hasil perhitungan koefisien reliabilitas tes 0,790 untuk instrumen tes 5 pilihan dan 0,679 untuk instrumen tes 3 pilihan.

## HASIL PENELITIAN DAN PEMBAHASAN

### Deskripsi Data

Berdasarkan jawaban siswa setelah mengerjakan tes untuk masing-masing kelompok didapat estimasi kemampuan siswa untuk masing-masing kelompok. Nilai rata-rata terendah pada kelompok CS3 yakni -0,008 dan tertinggi 0,057. Nilai minimum -1,979 pada kelompok PS5 dan maksimum 2,717 pada kelompok CS5. Bila dikelompokkan menjadi 3 kelompok yakni kelompok kemampuan rendah dari -2 sampai dengan -1, kemampuan sedang dari -1 sampai dengan 1, dan kemampuan tinggi dari 1 sampai dengan 3.

**Tabel 1.** Distribusi Estimasi Kemampuan Siswa

| Kemampuan | CS3 | CS5 | PS5 | RS5 |
|---|---|---|---|---|
| (-2,00) – (-1,01) | **40** | **38** | **61** | **53** |
| (-1,00) – 0,00 | 236 | 236 | 189 | 180 |
| 0,01 – 1,00 | 150 | 151 | 165 | 198 |
| 1,01 – 2,00 | **71** | **56** | **81** | **63** |
| 2,01 – 3,00 | **3** | **19** | **4** | **6** |
| Rerata | -0,008 | 0,030 | 0,036 | 0,057 |
| Minimum | -1,388 | -1,526 | -1,979 | -1,754 |
| Maksimum | 2,470 | 2,717 | 2,598 | 2,358 |

**Tabel 2.** Distribusi Tingkat Kemampuan Peserta

| Kemampuan | CS3 | CS5 | PS5 | RS5 |
|---|---|---|---|---|
| Rendah | 40 (8%) | 38 (7,6%) | 61 (12,2%) | 53 (10,6%) |
| Sedang | 386 (77,2%) | 387 (77,4%) | 354 (70,8%) | 378 (75,6%) |
| Tinggi | 74 (14,8%) | 75 (15,0%) | 85 (17,0%) | 69 (13,8%) |

Distribusi indeks kewajaran skor siswa dengan model penskoran *correct score* dengan tiga pilihan (CS3) dan penskoran *correct score* dengan lima pilihan (CS5) berturut-turut tercantum pada tabel 3. Distribusi indeks kewajaran skor siswa yang mengerjakan soal lima pilihan dengan model penskoran *punishment score* (PS5) dan penskoran *reward skor* (RS5) berturut-turut tercantum pada tabel 4.

**Tabel 3.** Distribusi Indeks Kewajaran Skor Kelompok

| Skor | Indeks Kewajaran | Proporsi CS3 | | Proporsi CS5 | |
|---|---|---|---|---|---|
| | | $\theta$ **rendah** | $\theta$ **tinggi** | $\theta$ **rendah** | $\theta$ **tinggi** |
| Tak Wajar | < (- 2,000)<br>(-2,000) – (-1,001)<br>-1,000 – 0,000 | 0,725 | 0,257 | 0,421 | 0,320 |
| Wajar | 0,001 – 1,000<br>1,001 – 2,000<br>> 2,000 | 0,275 | 0,743 | 0,579 | 0,680 |

**Tabel 4.** Distribusi Indeks Kewajaran Skor Kelompok PS5

| Skor | Indeks Kewajaran | Proporsi PS 5 | | Proporsi RS 5 | |
|---|---|---|---|---|---|
| | | $\theta$ **rendah** | $\theta$ **tinggi** | $\theta$ **rendah** | $\theta$ **tinggi** |
| Tak Wajar | < (- 2,000)<br>(-2,000) – (-1,001)<br>-1,000 – 0,000 | 0,457 | 0,153 | 0,453 | 0,145 |

| Skor | Indeks Kewajaran | Proporsi PS 5 | | Proporsi RS 5 | |
|---|---|---|---|---|---|
| | | θ rendah | θ tinggi | θ rendah | θ tinggi |
| Wajar | 0,001 – 1,000<br>1,001 – 2,000<br>> 2,000 | 0,525 | 0,847 | 0,547 | 0,855 |

**Pengujian Hipotesis**

***Pengujian hipotesis perbedaan proporsi skor wajar kelompok correct score yang mengerjakan soal tiga pilihan dan lima pilihan***

Pengujian hipotesis statistik penelitian dilakukan dengan menggunakan statistik z untuk pengujian statistik proporsi. Hasil uji menunjukkan pada kelompok kemampuan rendah mempunyai perbedaan proporsi 30,4%. Bila dilihat dari nilai z hitung = 2,716 > dari 2,33 maka perbedaan proporsi ini sangat signifikan dengan ukuran efek termasuk kategori sedang yakni ES = 0,63. Tetapi untuk kelompok kemampuan tinggi dengan perbedaan proporsi hanya 6,3% dan perbedaan ini tidak signifikan karena z hitung = 0,849 < 1,96.

**Tabel 5.** Hasil Pengujian Proporsi pada Model *Correct Score*

| Kemampuan | Kelompok | N | Proporsi (%) | Selisih proporsi | z hitung | z tabel 0,05 |
|---|---|---|---|---|---|---|
| Rendah | 3 pilihan | 40 | 27,5 | 30,4 | 2,716 | 1,65 |
| | 5 pilihan | 38 | 57,9 | | | |
| Tinggi | 3 pilihan | 74 | 74,3 | 6,3 | 0,849 | 1,96 |
| | 5 pilihan | 75 | 68,0 | | | |

***Pengujian hipotesis mengenai perbedaan proporsi skor wajar untuk siswa yang mengerjakan soal lima pilihan antara model penskoran CS dan PS, CS dan RS serta PS dan RS untuk masing-masing tingkat kemampuan siswa***

Berdasarkan tabel 5 dapat dilihat bahwa untuk kelompok siswa dengan kemampuan rendah perbedaan proporsi pasangan CS5 dan PS5, PS5 dan RS5, serta CS5 dan RS5 berturut-turut 5,4%; 2,2% dan 3,2% yang menghasilkan z hitung 0,53; 0,23 dan 0,30. Bila dibandingkan z tabel maka semua nilai z hitung tersebut kurang dari 1,96 dengan demikian perbedaan proporsi tersebut tidak signifikan. Untuk kelompok dengan tingkat kemampuan tinggi, pasangan CS5 dan PS5 maupun CS5 dan RS5 menghasilkan perbedaan proporsi berturut turut 16,7% dan 17,5% dengan nilai z hitung masing-masing 2,50 dan 2,47 sehingga dapat disimpulkan tolak $H_0$, dengan kata lain proporsi skor wajar kelompok PS5 dan RS5 lebih besar dibandingkan kelompok CS5, tetapi tidak ada perbedaan proporsi skor wajar antara kelompok PS5 dan RS5. Artinya untuk siswa dengan kemampuan tinggi, pada soal dengan lima pilihan penggunaan model penskoran *punishment score* dan *reward score* menghasilkan skor wajar yang lebih banyak dibandingkan penggunaan model penskoran *correct score*, dengan ukuran efek

(ES) sedang yakni ES = 0,46 dan ES = 0,36. Sementara itu penggunaan model penskoran *punishment score* menghasilkan proporsi skor wajar yang tidak berbeda dibandingkan dengan penggunaan model penskoran *reward score*. Dengan kata lain data tidak mendukung untuk menolak $H_{0.}$

Model penskoran yang digunakan dalam penskoran dan jumlah pilihan ganda pada soal tes pilihan ganda dapat menghasilkan proporsi skor wajar yang berbeda atau sama dengan tingkat kemampuan peserta. Hasil pengujian proporsi menunjukkan bahwa pada kelompok siswa yang dikoreksi dengan model penskoran *correct score*, pada tingkat kemampuan rendah porporsi skor wajar siswa yang mengerjakan soal lima pilihan berbeda secara signifikan dibandingkan siswa yang mengerjakan soal tiga pilihan. Proporsi skor wajar siswa yang mengerjakan soal tiga pilihan yang hanya 27,5%. Namun pada tingkat kemampuan tinggi, perbedaan tersebut tidak signifikan. Proporsi skor wajar siswa yang menjawab soal lima pilihan 68% sedangkan proporsi skor wajar siswa yang menjawab soal pilihan tiga pilihan 74,3%. Apabila dilihat dari tingkat kemampuan siswa, pada tingkat kemampuan tinggi menghasilkan skor wajar lebih banyak dibandingkan pada tingkat kemampuan rendah baik soal tiga pilihan maupun soal lima pilihan. Berdasarkan indeks kewajaran yang dihitung menggunakan teori respon butir menunjukkan bahwa skor wajar yang lebih banyak pada kelompok siswa yang mengerjakan soal dengan lima pilihan bila digunakan model penskoran *correct score*.

**Tabel 6.** Hasil Pengujian Proporsi Kelompok Lima Pilihan

| Kemampuan | Kelompok | N | Proporsi (%) | Selisih Proporsi (%) | z hitung | z tabel 0,05 |
|---|---|---|---|---|---|---|
| Rendah | *Correct Score (CS)* | 38 | 57,9 | CS – PS 5,4 | 0,53 | 1,96 |
| | *Reward Score (RS)* | 53 | 54,7 | CS – RS 3,2 | 0,30 | |
| | *Punishment Score (PS)* | 61 | 52,5 | PS – RS 2,2 | 0,23 | |
| Tinggi | *Correct Score (CS)* | 75 | 68,0 | CS – PS 16,7 | 2,50 | 1,65 |
| | *Reward Score (RS)* | 69 | 85,5 | CS – RS 17,5 | 2,47 | |
| | *Punishment Score (PS)* | 85 | 84,7 | RS5 – CS5 0,08 | 0,14 | 1,96 |

Bila dikaitkan dengan teori kemungkinan yang menyatakan makin banyak jumlah pilihan makin kecil peluang untuk menjawab benar bagi siswa yang mencoba untuk mengerjakan dengan menggunakan tebakan. Ternyata dapat ditunjukkan bahwa makin banyak jumlah pilihan skor yang diperoleh cenderung

makin wajar. Banyaknya skor wajar pada soal dengan lima pilihan untuk siswa dengan tingkat kemampuan rendah disebabkan karena tebakan yang dilakukan kebanyakan tebakan murni sehingga peluang menjawab salah lebih besar pada soal dengan lima pilihan dibandingkan soal tiga pilihan. Hasil ini sesuai dengan pendapat Thorndike (1997: 476) bahwa tebakan pada suatu tes memunculkan masalah serius pada tes yang memiliki sedikit pilihan jawaban dibandingkan tes dengan jumlah pilihan lebih banyak. Tebakan murni ini dilakukan siswa mengingat model penskoran yang digunakan adalah model penskoran *correct score* yang tidak memberikan sanksi terhadap jawaban yang salah. Sehingga skor yang diperoleh pada soal lima pilihan lebih wajar dibandingkan skor yang diperoleh pada soal tiga pilihan. Sedangkan pada tingkat kemampuan tinggi, kemungkinan besar siswa tidak menggunakan tebakan murni kecuali pada soal-soal yang sangat sulit sehingga skor yang diperoleh masih mencerminkan kemampuan peserta sesungguhnya.

Hasil pengujian pada siswa yang mengerjakan soal lima pilihan menunjukan bahwa pada tingkat kemampuan rendah, penerapan model penskoran tidak menghasilkan proporsi skor wajar yang berbeda. Jika dilihat proporsinya untuk model *correct score* 54,9%, pada model *punishment score* 52,5% dan model *reward score* 54,7%. Tetapi pada tingkat kemampuan tinggi, proporsi skor wajar pada model penskoran *punishment score* dan *reward score* lebih tinggi dibandingkan pada model *corrret score*. Pada model *correct score* proporsi skor wajar sebesar 68% sedangkan untuk *punishment score* dan *reward score* berturut-turut 84,7%, dan 85,5%. Sementara itu antara model penskoran *punishment score* dan *reward score* tidak terdapat perbedaan proporsi skor wajar, hal ini sesuai dengan konsep penskoran bahwa kedua model akan menghasilkan peringkat yang sama.

Hasil penelitian ini menunjukan bahwa efek penggunaan model penskoran sangat tergantung pada tingkat kemampuan peserta. Berkaitan hal ini Frary (1980: 78) mengatakan bahwa keputusan untuk menggunakan model penskoran tergantung pada beberapa faktor antara lain: 1) kemampuan peserta tes untuk mengikuti instruksi. Sebagian besar peserta tes tidak memahami intruksi yang kompleks kapan untuk menebak dan kapan untuk tidak menebak, 2) sumber daya yang digunakan untuk perhitungan rumus. Berkaitan dengan rumus yang digunakan menggunakan proposisi instruksi-instruksi yang lebih kompleks bagi peserta tas, langkah-langkah tambahan dalam penyekoran, dan tantangan untuk menginformasikan kepada pengguna skor serta efek dari rumus penskoran yang digunakan, 3) tingkat peningkatan karakteristik psikometrik dari skor tes (validitas dan reliabilitas). Penggunaan rumus penskoran tidak disarankan jika tidak ada peningkatan makna terhadap karakteristik psikometrik, 4) tingkat kecurangan dari peserta tes. Sebelum rumus penskoran digunakan secara berkelanjutan untuk suatu situasi tes, penyelidikan perlu dilakukan untuk menentukan apakah ada perilaku yang tidak sesuai harapan dalam tes, apakah

gagal menebak karena memang peserta tidak tahu jawaban atau gagal mencapai butir-butir terakhir karena keterlambatan mengerjakan tes.

Sebagai temuan hasil penelitian ini adalah ketidakwajaran skor yang terjadi pada penyelenggaraan tes dengan instrumen tes pilihan ganda dengan lebih banyak pilihan jawaban menghasilkan skor–skor yang lebih wajar dibandingkan instrumen tes pilihan ganda dengan lebih sedikit pilihan jawaban khususnya pada siswa dengan tingkat kemampuan rendah. Temuan ini didukung teori peluang bahwa frekuensi harapan jawaban benar hasil tebakan lebih kecil pada tes lima pilihan dibandingkan tes tiga pilihan. Dari sudut pandang psikometri, model penskoran *correct score* memunculkan perilaku *random guessing* karena model ini tidak mempertimbangkan pengetahuan parsial (*partial knowledge*) yang dimiliki peserta tes (Abu-Sayf, 1977: 853). Siswa dengan kemampuan tinggi pada umumnya menggunakan teknik *partial information* yakni kemampuan mengeliminasi pilihan jawaban yang pasti salah sehingga pilihan jawaban yang tersisa menjadi lebih sedikit termasuk jawaban yang benar dan peluang menjawab benar menjadi besar (Frary, 1980: 80). Dengan demikian proporsi skor wajar tidak berbeda antara kelompok yang menjawab tes dengan lima pilihan jawaban dan kelompok yang menjawab tes dengan tiga pilihan jawaban

Ditinjau dari model penskoran, pada umumnya skor siswa yang dikoreksi menggunakan model penskoran *punishment score* dan *reward score* menunjukkan hasil lebih akurat sesuai dengan kemampuannya dibandingkan penskoran *correct score* (Gronlund and Linn, 1990: 243). Dalam penelitian ini hanya berlaku untuk siswa dengan kemampuan tinggi yakni proporsi skor wajar pada model penskoran PS dan RS lebih besar dibandingkan proporsi skor wajar pada penskoran PS. Tetapi antara kedua penskoran PS dan RS menjaring skor wajar dengan proporsi yang tidak berbeda, dan kesimpulan ini tidak berlaku pada siswa dengan kemampuan rendah. Penelitian ini mendukung hasil penelitian Yuliatri yang menyimpulkan bahwa penskoran PS dan RS menghasilkan rata-rata fungsi informasi butir yang lebih tinggi dibandingkan penskoran CS yang menunjukkan ketepatan suatu parameter yang diukur. Penskoran PS dan RS mempunyai nilai fungsi informasi butir yang tidak berbeda (Sastrawijaya, 2005: 139). Sementara itu, siswa dengan kemampuan rendah baik ditinjau dari jumlah pilihan maupun dari model penskoran lebih banyak memiliki skor yang tak wajar dibandingkan dengan siswa kemampuan tinggi. Hal ini kemungkinan disebabkan siswa yang berkemampuan rendah tidak begitu memperhatikan model penskoran sehingga ketika tidak tahu jawaban atas butir yang dikerjakan tetap saja mereka menjawab dengan cara menebak.

## SIMPULAN

Melalui serangkaian kegiatan pengumpulan data dan analisis data dapat dirumuskan beberapa kesimpulan sebagai hasil penelitian sebagai berikut: ***Pertama***, pada tingkat kemampuan rendah, penggunaan model penskoran

*correct score* pada soal dengan lima pilihan menghasilkan skor yang lebih wajar dibandingkan soal dengan tiga pilihan jawaban. Tetapi pada tingkat kemampuan tinggi tidak terdapat perbedaan skor wajar diantara kedua kelompok tersebut. ***Kedua,*** pada tingkat kemampuan rendah penggunaan ketiga model penskoran pada soal dengan lima pilihan tidak menghasilkan perbedaan skor wajar. Tetapi pada tingkat kemampuan tinggi ada perbedaan skor wajar yang terjadi, khususnya model penskoran *punishment score* dan *reward score* menghasilkan skor yang lebih wajar dibandingkan model penskoran *correct score*. ***Ketiga,*** pada tingkat kemampuan tinggi pada soal dengan lima pilihan penggunaan model penskoran *punishment score* dan *reward score* menghasilkan proporsi skor wajar yang tidak berbeda antara kedua kelompok tersebut.

## DAFTAR PUSTAKA


Abu-Sayf, F. K. "A New Formula Score. (1977). "*Journal of Educational Psychological Measurement*, Vol. 4 (37\. Sage Publication. 1977, http://epm.sagepub. com/cgi/conten/abstract /37/4 /853.

Aiken, Lewis R. (1997). *Psychological Testing and Assessment*. Ninth Edition. Needham Heights, MA: Allyn & Bacon.

Allen, M. J., dan, W. M. Yen. (1979). *Introduction to Measurement Theory.* Monterey: Wardsworth, Inc.

Azwar, Saifuddin. (1987). *Tes Prestasi, Fungsi dan Pengembangan Pengukuran Prestasi Belajar*. Yogyakarta: Liberty.

Crocker, Linda dan James Algina. (1986). *Introduction to Classical and Modern Test Theory.* Florida: Holt, Rinchart, and Winston, Inc.

Ebel, Robert dan David Frisbie. (1991). *Essential of Education Measurement*. New Jersey: Prentice Hall, Inc.

Frary, Robert B. (2007). "The Effect of Misinformation, Partial Information, and Guessing on Expected Multiple-Choice Test Item Score." *Applied Psychological Measurement.* No. 4. Sage Publication. 1980. http://www.sagepub.com/ journals Reprints.nav.

_____. (2007). *Formula Scoring of Multiple-Choice Test. (Correction for Guessing)* Instructional Topics in Educational Measurement. National Council on Measurement in Education. University of Nebrasca-Lincoln. http://www. ncme.org/pubs/items/cfm.

Gronlund. E. N. dan Robert L. Linn. (1990). *Measurement and Evaluation in Teaching* 6th Edition. New York: Macmillan Publishing Company.

Hopkins, C. D. dan Anteso R. L. (1985). *Classroom Measurement and Evaluation*. Illinois: Peacock Publisher, Inc.

Hulin, C. I, Pritsz Drasgow, dan Charles K. Parsons. (1983). *Item Response Theory Applications to Psychological Measurement*. Illinois: Dow Jones-Irwin.

Lord, Frederic M. (1980). *Aplication of Item Response Theory to Practical Testing Problems* New Jersey: Lawrence Erlbaum Associates Publisher.

Naga, Dali Santun. (2012). *Teori Sekor Pada Pengukuran Mental*. Jakarta: PT Nagarani Citrayasa.

Nitko, Anthony J. (2001). *Educational Assesment of Students*. New Jersey: Prentice Hall, Inc.

Nunnaly, J. C. (1970). *Introduction to Psychological Measurement*. New York: McGraw-Hill Book Company.

Sastrawijaya, Yuliatri. (2005). ”Perbandingan Fungsi Informasi Butir Model Logistik Dua Parameter Ditinjau Dari Model Penskoran Pada Tes Pilihan Ganda Pada SMAN Jakarta Tahun 2004.” *Disertasi*, UNJ Jakarta.

Simanungkalit, Alfred. (1988). ”Hubungan Antara Sikap Terhadap Matematika, Ke-khawatiran Tes Terhadap Matematika, dan Locus of Control Terhadap Ma-tematika dengan Ketidakwajaran Jawaban Siswa Pada Tes Hasil Belajar Matematika Pada Sekolah Menengah Atas Di Wilayah DKI Jakarta.” *Disertasi*, IKIP Jakarta.

Thorndike, Robert L. (2005). *Measurement and Evaluation in Psychology and Education.* New Jersey: Pearson Education, Inc.

Wiersma, W., dan Stephen G. Jurs. (1990). *Educational Measurement and Testing,* Second Edition. Needham Heights, MA: Allyn and Bacon.